American Journal of Medical Genetics (Neuropsychiatric Genetics) 74:62–64 (1997)

Brief Research Communication

Bias in the Genomic Distribution of CAG and CTG Trinucleotide Repeats

Michael C. O'Donovan* and Carol Guy
Department of Psychological Medicine, University of Wales College of Medicine, Heath Park, Cardiff, UK

We investigated the hypothesis that the trinucleotide repeat CAG is disproportionately located in exons in genomic DNA by analyzing unbiased genomic sequences with the Gene Recognition and Analysis Internet Link program (http//avalon.epm.ornl.gov/). Forty percent of CAG/CTG repeats were predicted to lie within exons. This is significantly greater than would be expected by chance, and is also greater than we have observed for ATT/AAT repeats. Therefore, our data support the hypothesis. Furthermore, the data support the utility of a recently reported CAG/CTG PCR genomic screening set for identifying pathogenic expanded CAG/CTG repeats. *Am. J. Med. Genet. 74: 62–64, 1997.*

KEY WORDS: psychosis; expanded trinucleotide repeats; CAG; genomic distribution

INTRODUCTION

The trinucleotide repeat sequence CAG is the pathogenic mutation underlying five neurodegenerative diseases, and it is likely there are unidentified CAG repeat diseases, including SCA2 and other autosomal-dominant cerebellar ataxias [Trottier et al., 1995]. There is also strong evidence that expanded CAG or CTG repeats are involved in the pathogenesis of schizophrenia and bipolar disorder [O'Donovan et al., 1995, 1996; Lindblad et al., 1995; Morris et al., 1995], but unfortunately, the rapid identification of pathogenic expanded repeats in these diseases is hampered by difficulties in cloning large repeat sequences directly from genomic DNA [Gastier et al., 1996]. Recently, a PCR screening set for CAG repeats was identified [Gastier et al., 1996]. Although it is not yet complete, in principle, the set offers a systematic method for mapping pathogenic expanded CAG repeats. A major advantage of this screening set is that because the CAG repeats were derived from genomic DNA clones, no further delineation of intron-exon boundaries is required prior to the amplification of genomic DNA. This must, however, be balanced by the potential disadvantage that a proportion of the repeats will not be located within coding sequences of genes, and since all known pathogenic expanded CAG repeats are contained within exons, intronic and extragenic sequences may have a low probability of being involved in disease.

A recent survey of the distribution of trinucleotide repeat sequences where the number of repeats was greater than four produced the striking result that CAG sequences were apparently excluded from intronic DNA [Stallings, 1994]. However, this conclusion was based upon analysis of two sequences lodged in GenBank in which coding sequences are overrepresented. Indeed, it would be anticipated that this bias would be particularly marked for CAG repeat sequences, as there have been specific endeavors to clone them from expressed sequences following the discovery of expanded CAG repeat diseases [e.g., Li et al., 1993]. Nevertheless, the findings were specific to CAG but not to other repeats, and the magnitude of the effect seems unlikely to be explicable by bias. The systematic cloning of CAG repeats of five or more repeat units from genomic DNA without bias towards expressed sequences [Gastier et al., 1996] now offers the opportunity to reexamine this phenomenon.

We randomly retrieved 255 of the genomic sequences which were described by Gastier et al. [1996] as containing CAG repeats, using GCG software [Genetics Computer Group, 1994]. Sequences (accession codes available by request) were retrieved using the lookup command followed by fetch command. Coding sequences within each complete genomic sequence were then predicted using the Gene Recognition and Analysis Internet Link (GRAIL) [Xu et al., 1995], Version 1a. To yield an estimate of the expected number of similar-sized sequences which contain a repeat which would be predicted by GRAIL to be within exons, 157 genomic sequences containing five or more ATT/AAT repeats were also retrieved and analyzed the same way. ATT/AAT repeats were selected for this purpose because there is a large data base of sequences lodged in GenBank which were derived directly from genomic DNA and which contain this repeat [Gastier et al., 1995].

*Correspondence to: Michael O'Donovan, Department of Psychological Medicine, University of Wales College of Medicine, Heath Park, Cardiff CF4 4XN, UK.

Received 22 April 1996; Revised 29 July 1996

Repeats were described by GRAIL as being located in excellent, good, or marginal exons (Table I). The probability of sequences classified by GRAIL as excellent exons, good exons, and marginal exons representing true exons has been estimated at 100%, 69%, and 16%, respectively (GRAIL handbook). Marginal exons were therefore discounted from further analyses.

Assuming that all excellent and good exons are true exons, 40% of CAG/CTG repeats were predicted to lie within the coding sequence. This percentage is lower than that observed by Stallings [1994], but is considerably greater than would be predicted by chance, assuming that 10% of the genome is coding sequence ($X^2 = 233.6$, 1 df, $P < 10^{-6}$). The latter is a conservative assumption based upon a haploid genome size of about 3.3×10^9 bases, an average mRNA size of 2.2 kb, and about 150,000 genes [Lewin, 1994]. Under an even more conservative model in which only excellent exons represent true exons, and the proportion of coding sequence is doubled to 20% (postulating the existence of 300,000 genes), the proportion of CAG/CTG repeats predicted within coding sequence is still greater than expected by chance ($X^2 = 39.7$, 1 df, $P < 10^{-6}$).

CTG sequences were more likely than CAG sequences to lie within exons ($X^2 = 4.01$, 1 df, $P = 0.0452$), although this is not significant after correction for multiple testing. This finding is in disagreement with Stallings [1994], who observed that a greater proportion of CAG than CTG repeats lay in exons. However, in that study only a small number of CTG repeat sequences were identified, and therefore the result may have been due to sampling variance.

Only one ATT/AAT repeat was predicted to lie within an exon. This observation is consistent with previous findings [Stallings, 1994]. ATT/AAT repeats were significantly less likely than CAG/CTG repeats to lie within exons ($X^2 = 79.28$, $P < 10^{-6}$). The reason for the difference in the distributions of ATT/AAT and CAG/CTG repeats is unclear and cannot simply be accounted for by a systematic difference in the way GRAIL handles the two repetitive sequences [Y. Xu, GRAIL staff, personal communication], or by the fact that in one frame of six (UAA), the former repeat is a termination codon. As we do not know the minimum proportion of the genome that must be represented in exons, we are uncertain if our data represent a random distribution of ATT/AAT repeats within exons or a relative exclusion of this motif from the coding sequence.

TABLE I. Distribution of Repeats in Exon Sequences Predicted by GRAIL*

Repeat	Excellent	Good	Marginal
CAG	34	4	2
CTG	52	9	5
AAT	0	0	0
ATT	0	1	0

*Sequences of 255 genomic clones containing CAG/CTG repeats and 157 genomic clones containing ATT/AAT repeats were retrieved from GenBank and analyzed for coding sequences using GRAIL. GRAIL predicted the strand (e.g., CAG vs. CTG repeat) and classified exons as excellent, good, or marginal.

One hypothesis for the apparent overrepresentation of CAG repeats within exons is that the similarity of CAG to a consensus splicing sequence results in relative exclusion of the repeat from introns [Stallings, 1994]. However, as predicted by GRAIL, the proportions of exonic CAG and CTG repeats which are located within or adjacent to putative splicing sites are only 25% (7/34) and 17% (9/52), respectively. Therefore, it is unlikely that similarity to consensus splicing sequences can account for the dramatic excess of CAG/CTG repeats within exons.

The data presented here should be treated with caution, as they are derived from predictions made by computer algorithms, and there are several reasons why we may have *underestimated* the bias of CAG/CTG repeats towards exonic location. First, GRAIL has a false-positive rate of about 11% and a false-negative rate of about 18% (GRAIL handbook). Second, the false-negative rate may be even higher in this study, as the sequences analyzed were of a modest size (CAG/CTG, μ = 416 bp, SD = 121.05). Third, GRAIL predictions are restricted to translated sequences, and therefore CAG/CTG repeat sequences located in untranslated exons (e.g., the myotonic dystrophy repeat [Brook et al., 1992]) will be missed.

In conclusion, our analysis of genomic sequences supports the hypothesis that CAG repeats are overrepresented within coding sequence, as previously observed [Stallings, 1994], but not in a strand-specific fashion. Consequently, our data also support the usefulness of the published screening set [Gastier et al., 1996] as an efficient strategy for identifying the pathogenic CAG/CTG expanded trinucleotide repeats which appear to contribute to psychosis.

ACKNOWLEDGMENTS

This work was funded by the Welsh Scheme for the Development of Health and Social Research and the Medical Research Council. We are grateful for advice on the GRAIL program given by staff at the UK Human Genome Mapping Project Resource Center Computing Advisory Service.

REFERENCES

Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, Hunter K, Stanton VP, Thirion JP, Hudson T, Sohn R, Zemelman B, Snell RG, Rundle SA, Crow S, Davies J, Shelbourne P, Buxton J, Jones C, Juvonen V, Johnson K, Harper PS, Shaw DJ, Housman DE (1992): Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member Cell 68:799–808.

Gastier JM, Pulido JC, Sunden S, Brody T, Buetow KH, Murray JC, Weber JL, Hudson TJ, Sheffield VC, Duyk GM (1995): Survey of trinucleotide repeats in the human genome: Assessment of their utility as genetic markers Hum Mol Genet 4:1829–1836.

Gastier JM, Brody T, Pulido JC, Businga T, Sunden S, Hu X, Maitra S, Buetow KH, Murray JC, Sheffield VC, Boguski M, Duyk GM, Hudson TJ (1996): Development of a screening set for new $(CAG/CTG)_n$ dynamic mutations. Genomics 32:74–85.

Genetics Computer Group (1994): "Program Manual for the Wisconsin Package, Version 8." Madison, Wisconsin: USA 53711.

Lewin B (1994): "Genes V." Oxford: Oxford University Press, pp 657–676.

Li SH, McInnis MG, Margolis RL, Antonarakis SE, Ross CA (1993): Novel triplet repeat containing genes in human brain: Cloning, expression, and length of polymorphism. Genomics 16:572–579.

Lindblad K, Nylander PO, De Bruyn A, Sourey D, Zander C, Engstrom C, Holmgren G, Hudson T, Chotai J, Mendlewicz J, Van Broeckhoven C, Schalling M, Adolfsson R (1995): Detection of expanded CAG repeats in bipolar affective disorder using the repeat expansion detection (RED) method. Neurobiol Dis 2:55–62.

Morris AG, Gaitonde E, McKenna PJ, Mollon JD, Hunt DM (1995): CAG repeat expansions and schizophrenia: Association with disease in females and with early age-at-onset. Hum Mol Genet 4: 1957–1961.

O'Donovan MC, Guy C, Craddock N, Murphy KC, Cardno AG, Jones LA, Owen MJ, McGuffin P (1995): Expanded CAG repeats in schizophrenia and bipolar disorder. Nat Genet 10:380–381.

O'Donovan MC, Guy C, Craddock N, Bowen T, McKeon P, Macedo A, Maier W, Wildenauer D, Aschauer HN, Sorbi S, Feldman E, Mynett-Johnson L, Claffey E, Nacmias B, Valente J, Dourado A, Grassi E, Lenzinger E, Heiden AM, Moorhead S, Harrison D, Williams J, McGuffin P, Owen MJ (1996): Confirmation of association between expanded CAG/CTG repeats and both schizophrenia and bipolar disorder. Psychol Med (in press).

Stallings RL (1994): Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: Implications for human genetic diseases. Genomics 21:116–121.

Trottier Y, Lutz Y, Stevanin G, Imbert G, Devys D, Cancel G, Saudou F, Weber C, David G, Tora L, Agid Y, Brice A, Mandel J-L (1995): Polyglutamine expansion as a pathological epitope in Huntington's disease and four dominant cerebellar ataxias. Nature 378:403–406.

Xu Y, Mural R, Uberbacher EC (1995): Correcting sequencing errors in DNA coding regions using dynamic programming. Comput Appl Biosci 11:117–124.